Original Paper

Title: Testing out suicide risk prediction algorithms using phone measurements with patients in acute mental health settings: a feasibility study

Abstract

Background: Digital phenotyping and machine learning are nowadays being used to augment or even replace traditional analytic procedures in many domains, including health care. Given the heavy reliance on smartphones and mobile devices around the world, this readily available source of data is an important and highly underutilized source that has the potential to improve mental health risk prediction and prevention and advance global mental health.

Objective: To apply machine learning in an acute mental health setting for suicide risk prediction. This study is using a nascent approach, adding to existing knowledge by using data collected through a smartphone in place of clinical data which has typically been collected from health care records.

Methods: We created a smartphone application called Strength Within Me (SWiM) that was linked to Fitbit, Apple Health kit, and Facebook, to collect salient clinical information such as sleep behavior and mood, step frequency and count, and engagement patterns with the phone from a cohort of acute mental health inpatients (n=66). In addition, clinical research interviews were used to assess mood, sleep, and suicide risk. Multiple machine learning algorithms were tested to determine best fit.

Results: K-Nearest Neighbors (k=2) with uniform weighting and Euclidean distance metric emerged as the most promising algorithm, with 68% average accuracy (averaged over 10,000 simulations of splitting the training and testing data via 10-fold cross validation) and average

AUC of 0.65. We applied a 5x2cv combined F test to test model performance of KNN against baseline classifier that guesses training majority, random forest, and others and achieved F statistics of 10.7 (p-value .0087), 17.6 respectively (p-value .0027), rejecting null of performance being the same. We have therefore taken the first steps in prototyping a system that could continuously and accurately assess risk of suicide via mobile devices.

Conclusions: Sensing for suicidality is an under-addressed area of research to which this paper makes a useful contribution. This is part of the first generation of studies to suggest it is feasible to utilize smartphone generated user input and passive sensor data/digital phenotyping to generate a risk algorithm among inpatients at suicide risk. The model reveals fair concordance between phone-derived and research generated clinical data and with iterative development has potential for accurate discriminant risk prediction. However, while full automation and independence of clinical judgement or input would be a worthy development for those individuals who are less likely to access specialist mental health services, and for providing a timely response in a crisis situation, the ethical and legal implications of such advances in the field of psychiatry need to be acknowledged.

Keywords:

Suicide prevention; suicidal ideation; suicide risk; smartphone; machine learning approaches; classification and regression trees; nearest neighbor algorithm; machine learning algorithms; digital phenotyping

Introduction

Limitations in scalability, accuracy and consistency with respect to traditional methods of predicting suicidal behavior have been recognized throughout literature and meta-analyses [1-

5]. Suicidality has been defined as any suicide-related behaviour, including completing or attempting suicide (intent), suicidal ideation (thoughts) or communications [6]. Not everybody who experiences suicidal ideation attempts suicide; but suicidal thoughts have been shown to be linked to higher risk of death by suicide [7]. While some people communicate their suicidal thoughts and/or plans to friends and family prior to suicide, others do not disclose their intent [8-10]. Also, some individuals might not seek help during a time of crisis due to various perceived constraints, including fear of stigma or disclosure, lack of time, access to services, preference for informal help [11]. Our ability to predict suicide is limited by our understanding of suicidal thoughts and their nature [12].

Advances in smartphones and connected sensors (wearables) have opened new possibilities for real time, context related monitoring of suicidal thoughts and suicidal risk [13], for example: ecological momentary assessments (EMA, [14]) that allow self-report of suicidal thoughts as they occur in an individual's day to day life/naturalistic setting [15], and digital phenotyping that enables access to real time classification and quantification of human behaviour [16-18]. The use of computational data-driven methodologies that use social media to understand health related issues (infodemiology/infoveillance, [19-20]) and data mining techniques (artificial intelligence/machine learning algorithms, [21]) provide additional potential in expanding our understanding of people's thoughts, feelings, behaviour, etc. and improving monitoring of suicide risk in real time. Although in its infancy, new research exploring suicidal ideation has shown that social media (e.g. Twitter, Facebook) could potentially be used as a suicide prevention tool [10, 22-26]. One study, for example, demonstrated the utility of social media blog post analysis in classifying individuals with high suicide risk in China [27]. Some research indicates that, by analysing certain patterns of smartphone use, changes in mental health symptoms could be identified [28].

Whereas standardised clinical tools can help towards classifying factors that contribute to suicide risk and understand biological markers related to suicide (trait analyses), computer science/machine learning can provide additional and timely tools to understand suicide thought linguistic markers (state analyses) [29]. New statistical methods have been proposed and tested to achieve more accurate predictions of risk, for example support vector machines, deep neural

nets, random forests [13]. Evidence suggests that these methods, especially elastic-net, perform better compared to traditional logistic regression techniques [30]. There is a shift towards developing more personalised risk profiles and using decision tree techniques exploring hundreds at predictors rather than a few clinically relevant risk factors [31]. Modern machine learning techniques are better placed to identify complex relationships between large datasets and suicide risk [13].

Early evidence generated by a pilot study using data from 144 patients with mood disorders suggests that machine learning algorithms using prior clinical data were successful in distinguishing between people that attempt suicide and those who do not, with a prediction accuracy between 65-72% [32].

While there has been a growing body of research seeking to augment or advance traditional methods with the aid of machine learning in clinical psychiatry [2, 4, 30, 33-38], the majority of studies rely on applying algorithms that learn from clinical data such as health care/electronic medical records, unstructured notes by providers and caretakers or some other data carefully gathered by health care professionals.

In this feasibility study, we aim to add to existing knowledge by using a nascent approach combining clinical data with proxy risk active and passive data collected from mobile devices to develop our algorithm. We have developed a software platform to enable us to collect data about inpatients in acute mental health settings via our own mobile application, SWiM ('Strength within Me'), a smartphone (iPhone), a wrist wearable (Fitbit), and from questionnaires administered by the research team. Active risk data - patient facing user interface (UI) modules (e.g. journaling, safety plan, mood meter) and passive risk data - data collected behind the scenes not requiring direct interaction from the patient (e.g. sleep monitoring) were collected. This information was then used to construct and train the machine learning algorithms seeking to produce a risk score that deduces the likelihood of suicide. We used the risk level from the Columbia-Suicide Severity Rating Scale (C-SSRS) [39] assessed by mental health researchers as our standard classification target. C-SSRS is currently considered the 'gold standard' approach for the measurement of suicidal ideation and behaviour in clinical trials [40]. Previous research has confirmed the validity of the scale and its prediction accuracy

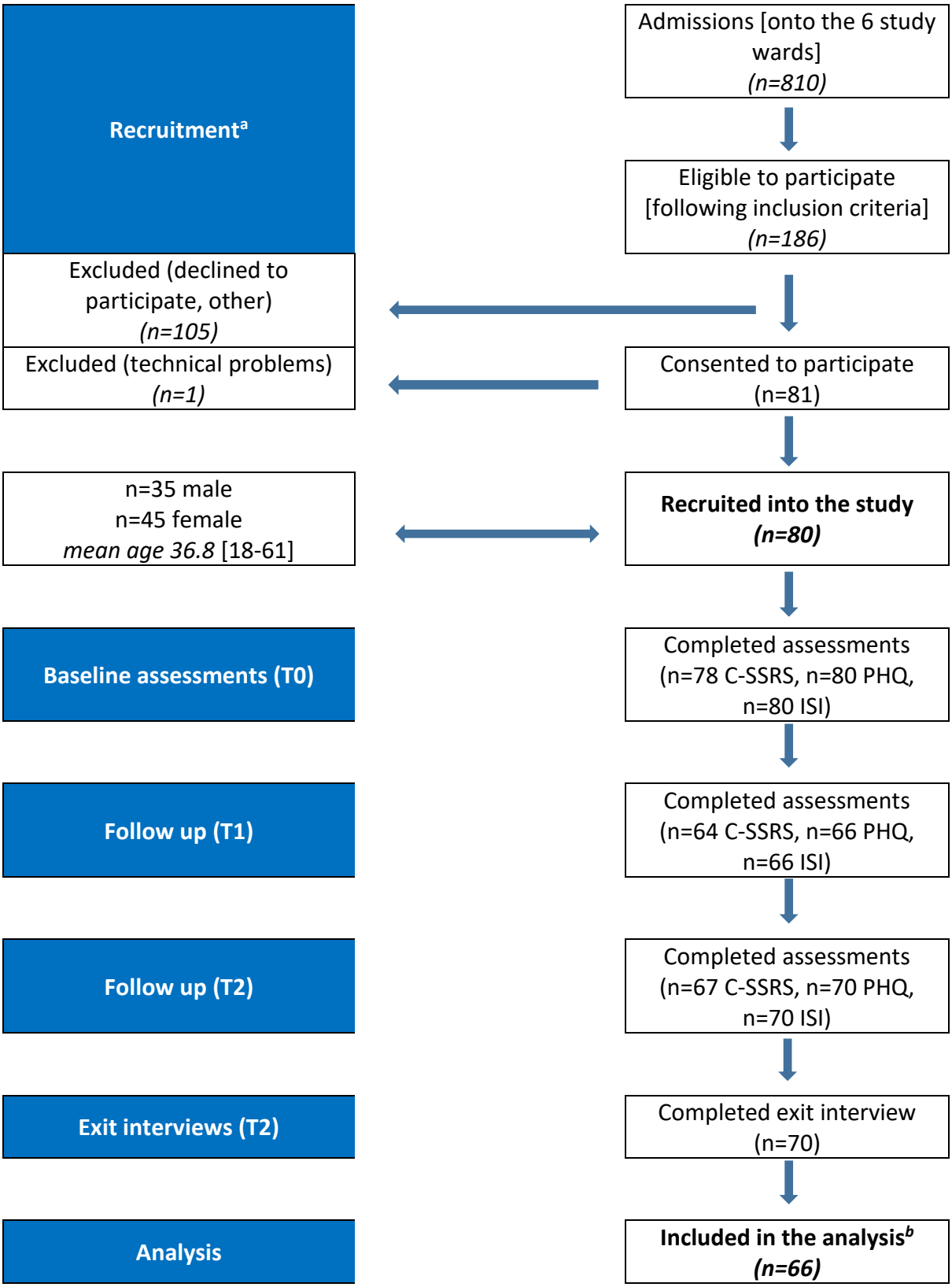
for short term risk of suicidal behaviour in clinical and research settings. Studies have demonstrated that individuals who meet the criteria of 'high risk' following administration of C-SSRS are almost four times more likely to attempt suicide within 24 months [39]. The C-SSRS was then compared with data from proxies for risk factors [41] such sleep quality and emotional health collected via Fitbit (Fitbit, Inc.) trackers and the SWiM app that patients interacted with for a week during their admission. This is our first target as to beat the standard we believe that we must be able to meet or approximate it for future development of our own metric.

Methods

Participants and Clinical Setting

In this Phase 1 feasibility study we have collected data from service users admitted onto 6 acute adult mental health wards within an NHS trust in the North West of England, UK. Service users who had been admitted onto a ward within the last 7-10 days were assessed by nursing staff to determine study eligibility. Following informed consent, participants were given a study iPhone and Fitbit to enable use of the SWiM app and monitor their sleep and daily activity for up to 7 days. Participants were then involved in 3 interviews at 3 different time points to complete a battery of assessments, including the Columbia Suicidal Severity Rating Scale, C-SSRS) [39], looking at suicidal thoughts and behavior. The interviews were completed by two experienced researchers who were trained to administer the clinical assessments. If suicidal risk was highlighted during the interview, nursing staff were informed and an agreed protocol was followed to ensure safety. Participants were given vouchers following completion of assessments. In total, 80 patients of 186 eligible consented to participate and 66 were included in the analysis based on completion of at least two follow up clinical assessments. This represents a 43% response rate and 83% completion rate. For a breakdown of participants, see *Table 1.*

Table1. SWiM study flow diagram (adapted from CONSORT, 2010).



[a] Timeframe for recruitment: Jan-Nov 2018

[b] Based on completed C-SSRS at follow up assessment (T2)

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All procedures involving human subjects/patients were approved by [HRA, REC 17/NW/0173, IRAS 214244]. Written informed consent was obtained from all inpatients.

Overview of Participant Data Fed into the Modelling Process (Table 2)

On a high level, the data are segmented into: (1) data entered by the user into the SWiM app; (2) data collected passively by the SWiM app, the Fitbit wearable, and the Apple Health app; (3) data directly gathered by our researchers; and (4) social interaction data for those who gave permission. We gathered a total of 173 variables - a mix of raw data such as counts of number of journal entries and derived values/features that involve summary statistics, or other variations of the data (e.g. adding up minute wise sleep records to get total sleep time or number of interruptions in sleep). Social interaction/Facebook data was excluded from the analysis, due to low response rate, i.e. 8/80 participants gave permission/had access to Facebook.

User-inputted data included participant mood, free form journal entries, steps for personal safety plans, and custom reminders they could have set for themselves. From these entries we collected descriptive statistics such as average mood reported (Likert Scale 1-5), average character limit, maxima, minima, and raw counts. A particular derived variable of interest from journaling was the average sentiment derived for each journal entry. This sentiment (ranging from -1 for negative to +1 for positive) was calculated via a third-party model known as VADER (Valence Aware Dictionary and Sentiment Reasoner) [42] that is catered to sentiments expressed in social media but has proven itself in other domains. The idea behind using this model was to get a proxy for the indication of feelings by users as they write and reflect.

Data collected by the research team included socio-demographic information, such as age and gender, and clinical assessment data. The key information that we used in the modeling was the researchers' assessment of the patient through the C-SSRS, which was assessed a maximum

of three times (patient entry, one or more follow-ups during their hospital stay, and exit. All 80 of consenting users were at risk upon entry to the ward (when first test was done), so at this point no machine learning/prediction to be done. The initial thought was to compare results against an intermediate survey result and exit survey result, consider change in risk, but we did not have enough exit surveys for two different time period comparisons. 66/80 had at least a second survey where risk level was reassessed and that was the population included for prediction. There was a 3-7 day wait from the first assessment to the second assessment. Finally, we included passive data gathered via the phone and the Fitbit wearable, such as details about a user's step frequency and count from Apple's Health kit application, minute-level sleep data from Fitbit, and engagement patterns with the phone (e.g. number of logins to the SWiM app, times a certain section visited, etc.) Levels of engagement with study data are presented in Table 3.

Table 2. SWiM study data.

Data Source	Examples of Variables Collected	Examples of Raw Data	Examples of Derived Data
Facebook			
	Stats of Facebook activity, post activity	Number Posts: 5, Number likes total:100	Average Likes/Post:20
User input			
	Journal, Mood, Reminders, Safety Plan Steps	Journal Entry: "Last night was horrible. I couldn't sleep at all with the noise."	Sentiment: -0.8 Word Count: 12
Clinical team			
	Demographics, C-SSRS Responses	Age: 35, C-SSRS Risk Overall: Moderate	C-SSRS Risk Binary: 1

Passive sensor data			
	Sleep, Steps, Interactions	{"dateTime": "23:10:00", "value": "awake"}, {"dateTime":"23:11:00", "value": "asleep"}	Sleep Latency: 1 minute Average time asleep: 5 hours

Table 3. Engagement rate across active and passive data in the study.

Data source	Rate
Step related features (FitBit and iPhone)	
	40% (26/66)
Journal entries (self-documented via SWiM app)	
	68% (45/66)
Mood entries (self-reported via SWiM app)	
	80% (53/66)
Phone activity (data usage)	
	100%
Sleep (FitBit)	90% (59/66)

Modeling

Machine Learning Setup and Data Analysis in Our Clinical Setting

As a first step towards developing an algorithmic risk score that is valid in predicting suicide risk, we framed the problem as a supervised, binary classification problem in which users were categorized in terms of levels of risk of "low risk" vs "high risk" using the information specified above. These "low risk" and "high risk" labels were derived from the overall C-SSRS risk scores obtained after asking participants a range of questions on previous attempts, ideation, etc. Usually the three categories are "low", "moderate", and "high", but we grouped "moderate"

and "high" for the sake of tractability from a modeling perspective. From a machine learning perspective, this aids in what is commonly referred to as the "class imbalance" problem [43], where certain categories have relatively few labels to their other counterparts. This makes it statistically more difficult to identify and these categories as models are inclined to achieve high scores by predicting the most common class; we turned a distribution of 36 low, 5 moderate, and 25 high to 36 low and 30 high. Choosing a binary case was helpful in dealing with the class imbalance issue, as models are data dependent in terms of volume (i.e. the more examples, the better job they do in learning). This is especially critical when we take into account the limitations in our data; because in order to fairly judge the model performance, we must partition the data (a test set and training set via k-fold cross-validation [44]) to assess how well the model can predict risk on "new users" given what it is learned from "old users" [45]. From a risk-app perspective, although it would be ideal to put users on a continuum of risk levels, it is critical to first assess the feasibility of identifying users at discrete thresholds as well as seeing the degree to which we can match a current standard in risk assessment.

Our "low risk" and "high risk" categories were mapped to binary outputs of 0 or 1. Some features derived from a user's journal entry are the word length and sentiment score (ranging from negative with -1 to positive +1) (for further info, please refer to the source model from which this is derived [42]). To account for the time dependency in the data (multiple journal entries across multiple days for example), a majority of the features engineered were done so in a "summary statistics" fashion (mean, median, variance, etc.) For example, the average journal word count per day over the user's total number of entries were used to summarize one aspect of a user's journaling behavior over their time with the app.

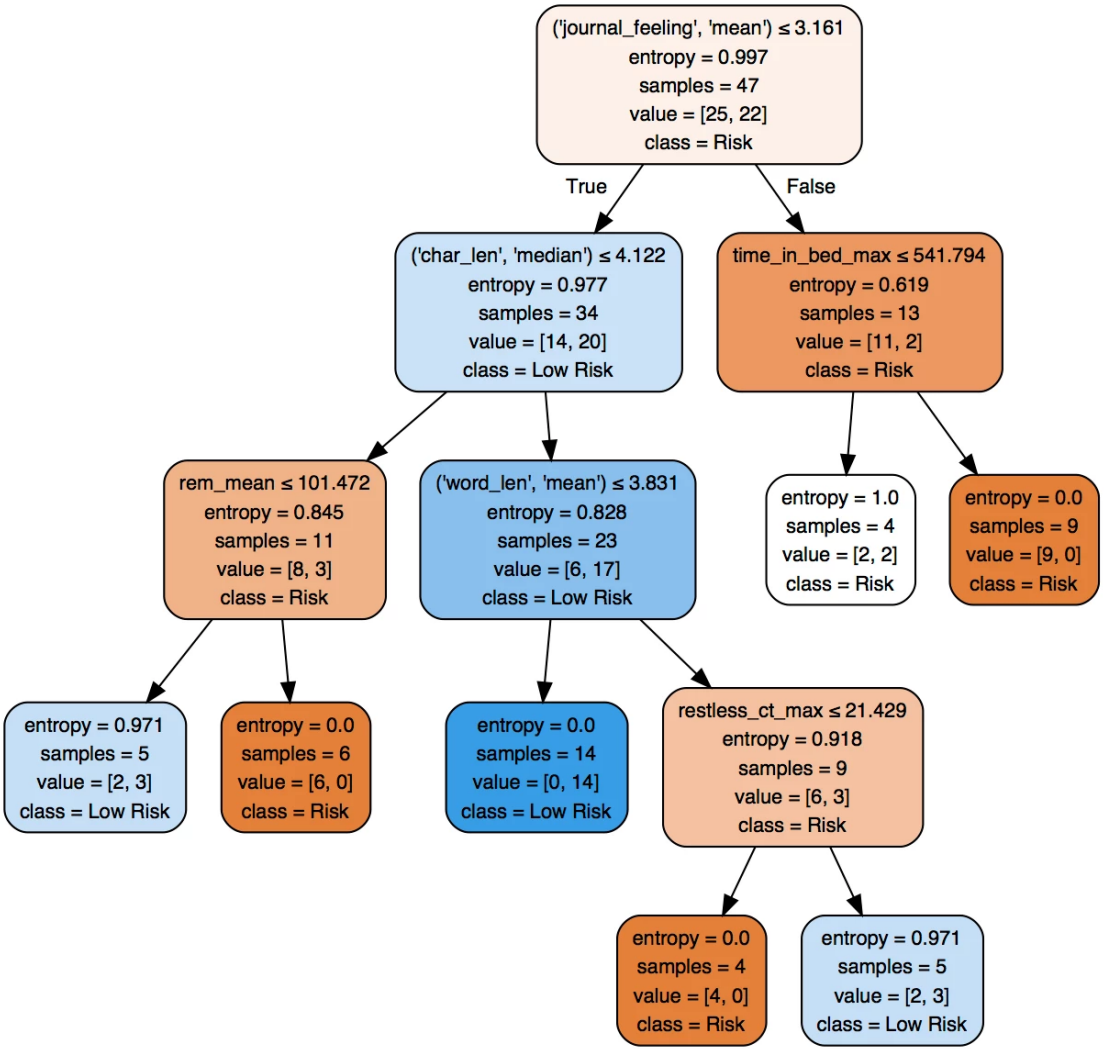
We curated 172 features formulated from categories of sleep data, journal entries, data usage, mood, and app activity statistics. For more information, a csv including full list of features incorporated into modeling (besides 'uid' which is user id to anonymize yet identify patient) is included in *Multimedia Appendix 1*. The 172 features were projected down to a 5-dimensional space by PCA. This sample is to help give insight for replicability. Any feature that has a summary statistic attached such as "mean" or "std" was done over the course of the 3-5 days

before the second assessment. Categorical features such as gender were mapped to numerical (in this case binary) outputs for the algorithm to consume.

This is typically considered a relatively high number of features relative to the amount of possible supporting data points/number of users recorded. In order to provide a more suitable set from which a machine learning algorithm may distinguish a signal for risk, we turned to feature selection and dimensionality reduction techniques. Our aim was to cut down to a smaller set of features that may also be interpretable and grounded in clinical knowledge of risk factors. We therefore opted for Principal Component Analysis (PCA) [46] as our dimension reduction technique and used Random Forests [47-49] to help in terms of feature selection as well to check the reliability of our reduction. Algorithms such as Support Vector Machines [50] are designed in such a manner as to overcome dimensionality issues, but they were experimentally confirmed to be unsuited to the task due to the size of the data.

For our study, the Random Forest model was composed of 25 Decision Trees. We took a look at the top 30 of the 170 original features and found that journal related features such as average feeling, cell activity such as the variation in user's data usage, sleep related features such as average sleep efficiency (time spent sleeping/total time spent in bed) and other natural indicators mostly known to clinical psychology as markers of risk. For an example of a decision tree formed for our data, see *Figure 1.*

Figure 1. Example of a decision tree formed for the SWiM study data[a].



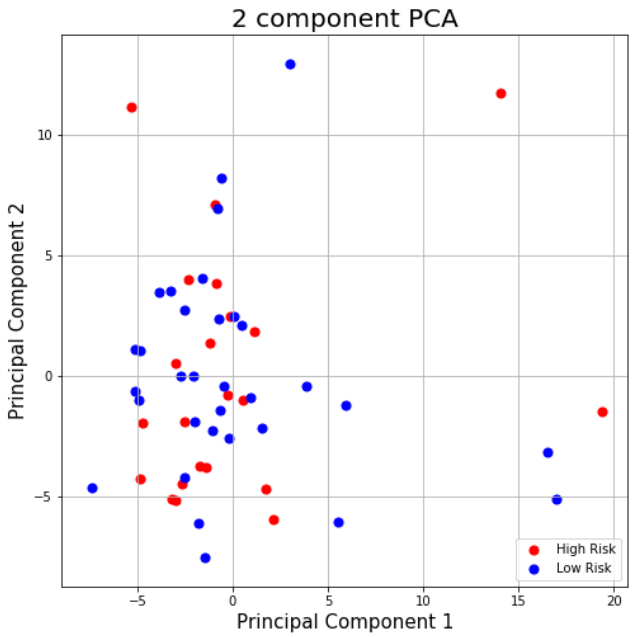
[a] The tree is read similarly to a flow chart in a top-down, left-right fashion. For example, at the top, we start with an entropy of .997 (entropy of 1 means complete uncertainty with 0 as certainty) [19-21] as we have 25 people in the low-risk category and 22 in the high-risk category. We then look at their average journal feeling, and if it is less than 3.161, we go to the left node with a sub-group of 34 people, otherwise the right node with a sub-group of 13 people. Following the right node, we now have a subgroup of 13 people with an average journal feeling greater than 3.161. Based on this characteristic alone, we reduce entropy to .619 (we are more certain of our group) and have 11 users correctly identified as high-risk, but 2 low-risk users misclassified as high-risk. Aiming to reduce misclassification, we again split by the average amount of time the user has spent in bed. If they have greater than 541 minutes/9 hours spent in bed in a day, a subgroup of 9 out of 9 people is correctly identified as high-risk. However, we see that for less than 9 hours, we also predict high-risk and have complete uncertainty (entropy

1), as the sub-group of 4 people are evenly divided amongst the classes. Once we reach one of these leaves/terminal nodes, we can read the decision process used to get there. For example, for the right-most leaf with 9 samples we discussed, users with an average journal feeling greater than 3.161, who also spend greater than 9 hours in bed, are identified as high-risk users. Similar interpretations can be made for the other 6 terminal nodes. Worth noting is that the features are ordered top-bottom in terms of ability to split classes and reduce entropy; by this criterion, we see journal feeling as the "most important" feature, time in bed as second, and so on.

Principal Component Analysis (PCA)

For dimensionality reduction and to guard against overfitting, we turned to Principal Component Analysis (PCA). On a high level, PCA groups together features that are correlated with one another into new features (principal components) that hold the most signal in terms of variation in the data [45]. The idea is that features that explain a high level of variability found in the data produce most of the signal needed to distinguish categories and discarding the rest of the features minimally loses signal but greatly reduces noise. Formally, PCA is an orthogonal linear transformation that maps the data to a new coordinate system such that a bulk of the variance of the projection is covered by the first k components, where k < total number of original features, and components are linear combinations of the originals. Another important characteristic of PCA is that it is not optimized for class separability and may be considered as an unsupervised model. This is critical as we aim to achieve generalizations outside of the data at hand and we do not want to over-fit our final model. To help give a visualization of the PCA transformation on our data, an example of a 2-dimensional/2 components PCA is given in Figure 2.

Figure2. A diagram of Principal Component Analysis[q].



[a] A high-dimensional dataset has been "flattened" to a 2-dimensional space where the new axes correspond to the principal components (they point in the direction of largest variance of the data).

After looking at the variance captured up to 100 possible components, we settled for the first 5 components as they accounted for 55% of the variance. Our first 5 components are described in *Table 4* below along with the themes/patterns identified after reviewing which features were grouped. We were assured that these components made sense in terms of clinical knowledge of how sleep quality, mood, activeness, and other characteristics are indicators to mental health [51-54]. Moreover, the top 30 features of our feature selection from Random Forest strongly overlapped with these features and so we were further assured in terms of potential predictive power.

Table 4. PCA components and patterns.

Component	Description	Themes/ Patterns
First Component		

	Max Efficiency, Average Efficiency, Median Efficiency, Max Time in Bed, Number Sleep Recordings	Ability to Sleep/Sleep Quality
Second Component		
	Number packets sent, Number times wifi sent, Number times cellular sent, Number times journal entered	User App Activity/ Data Presence
Third Component		
	SD Sleep Start, Median Journal Feeling, Max Sleep Start, Max Journal Feeling, Minutes in Bed, Min Journal Feeling	Feeling vs Sleep Activity
Fourth Component		
	Median Char Length, Median Word Length, Median Journal Feeling, SD Rest Duration, Max Rest Duration	Journal Input vs Resting Variability
Fifth Component		
	Median Sentiment, SD Number of Awakenings During Sleep, Number of Awakenings During Sleep, Minimum Sentiment	Sleep quality and Reflection Tone

Results

We tested a series of algorithms we thought would be best suited to predicting levels of risk from a theoretical perspective. Often referred to as the "bias-variance tradeoff" [45, 55-56], there is often the case with model selection that the best model should not be too simplistic such that its crude predictions miss a bulk of the cases, nor should it be overly complex such that its high sensitivity perfectly fits the data, but fails to generalize to new, unseen data. This

principle along with other individual algorithm properties helped guide the experimentation. As discussed within literature [57], increasing the complexity and flexibility of a model tends to allow it to understand more nuanced relations but at the cost of being overly sensitive to noise within data and over-fitting. Hence not only were models of varying complexities chosen to compare from linear models like logistic regression to non-parametric models such as knn, but the parameters within each model were also tuned by choosing number of neighbors and reducing dimensionality through PCA.

Just as important to mention is that these models are selected and judged based off of various metrics that aim to capture the objective for which the model is needed. Certain metrics also have advantages over other depending on the imbalance of classes, nature of the data (categorical or numerical), and other factors. Since we had a nearly balanced data set and this was a feasibility study, we opted for the simplest to understand metric of accuracy where we measured the number of correctly predicted observations over the total number. As a baseline, we looked at the simplest heuristic of predicting the majority class of low-risk users. This produced an average accuracy of 53%.

Random Forest was tested as it is generally agreed upon as a strong "out-of-the-box" model that does well on various datasets in different contexts, as well as having interpretability through the feature importance it can help provide [47,49]. Logistic Regression was another model considered due to the log-odds interpretability for the coefficients to each of the features (usually referred to as explanatory variables in explanatory contexts) and natural fit to classification problems [45]. Support Vector Machines (SVMs) [45, 50] were also tested as they have the design of naturally combating the "curse of dimensionality" through the transformations they do to the data ("kernel trick"). SVMs are also rather sophisticated models that tend to produce near state of the art results (barring neural networks which at the time of writing are highly data-hungry, and not necessarily interpretable). Finally, we considered the K-Nearest Neighbors algorithm which is often sought due to simplicity as well as the natural heuristic of classifying based off of how "close" observations are to one another [58].

To test performance, we performed k-fold cross validation with k=10. This means that we randomly partitioned the data into 10 pieces (folds) and used 9 of them to train the model and

one as an "unseen" piece (fold) to test on. This was done such that each of the 10 folds was used as the "unseen"/testing data at a given iteration. The idea was to get at the expected accuracy of a model when exposed to new data by simulating variations of data seen to unseen data. We repeated this process 10,000 times in order to get a more stable estimate as there are many ways to partition this data into 10 folds. *Table 5* summarizes the results.

Table 5. The average cross-validation accuracy along with the standard deviation of the accuracy observed for the various folds. K-Nearest Neighbors.

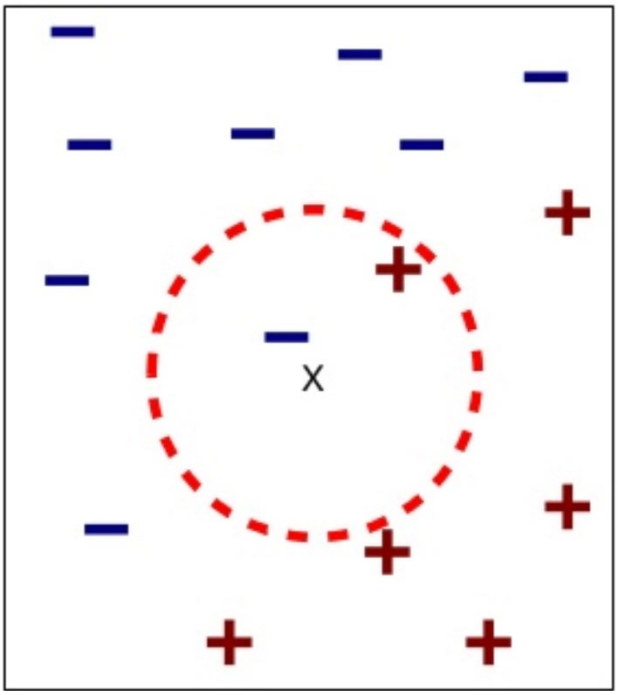
Algorithm	10-fold CV Avg Accuracy (10k iterations)	Standard Deviation	Comments
K-Nearest Neighbors (k=2) + PCA (n=5)			
	0.68	0.12	Best performance, k=2 seemed natural and worked best up to 10
Random Forest (k=25)+ PCA (n=5)			
	0.60	0.13	Non-linear helps, too many trees did not, PCA reduced deviation
Random Forest on raw features (k=25)			

	0.60	0.15	Non-linear helps, too many trees did not
SVM (degree 2 polynomial kernel)			
	0.57	0.10	Likely Overfit, Base Guessing
Logistic Regression + PCA (n=5)			
	0.59	0.14	Removed correlation due to PCA + prevent overfitting
Logistic Regression on raw features			
	0.55	0.16	Likely Overfit, Base Guessing
Baseline: Guessing Majority From Training Fold			
	0.53	.20	Baseline to Beat

Logistic Regression failed to do much better than baseline. With the raw features it performed poorly likely due to overfitting and high collinearity between some features (e.g. median sleep time and mean sleep time). We removed most of this through PCA and performed slightly better on average at 59%, but the standard deviation of 14% was worrying, given its below baseline lower end (worse than majority guessing). Similarly, SVM failed to do much better, and of the different "kernels" we present the polynomial degree two kernel as it performed best out of other variations (higher order polynomials, radial basis kernels, linear). We defer explanations of these kernels to literature. Random Forest did better than either of the other

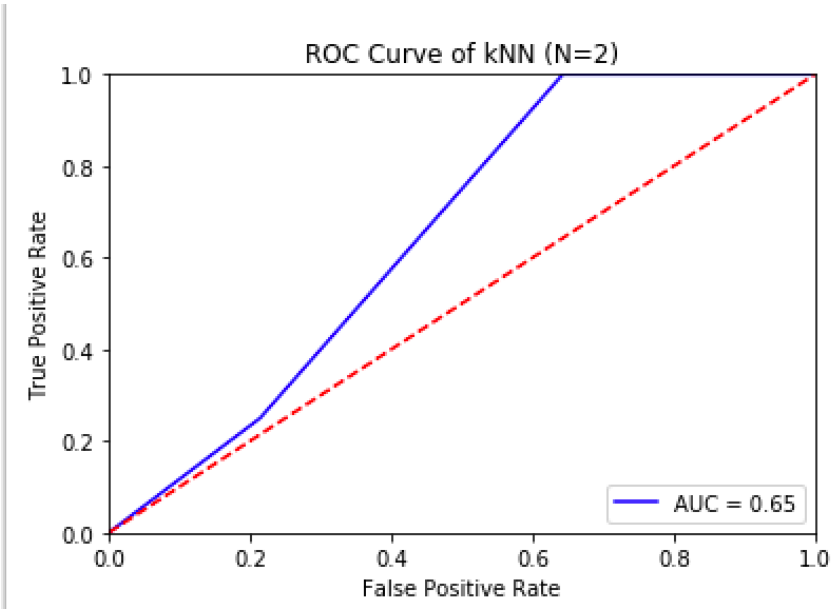
two mentioned algorithms, but worst-case folds still fell below baseline. The most promising was the K-Nearest Neighbors algorithm with k=2 with the 5 principal components discussed earlier as features with not only an average accuracy of 68% (averaged over 10,000 simulations of splitting the training and testing data via 10-fold cross validation) but with a standard deviation of 12% that puts its worst performance just above baseline at 56% and upper limit of 82% (*Figure 3*). In terms of false positive and true positive rates, the model achieved an average AUC of .65 (*Figure 4*). We applied a 5x2cv combined F test to test model performance of KNN against baseline classifier that guesses training majority, random forest, and others and achieved F statistics of 10.7 (p-value .0087), 17.6 respectively (p-value .0027), rejecting null of performance being the same [59-60].

Figure 3. Example of Nearest Neighbors with k=2 with data in two dimensions[a].



[a] Here the new test point is x and has 1 "minus neighbor" and "1 plus neighbor" as its two closest neighbors. Since the minus neighbor is closer, the new point x will be classified as minus.

Figure 4. ROC curve for KNN.



Due to the promising results of the algorithm, we explain it to readers unfamiliar with it. The K-Nearest Neighbors algorithm essentially follows the saying of "birds of the same feather flocking together". That is to say, the way prediction is done via this algorithm is that for a new test point, the distance (usually the well-known Euclidean distance) is computed between the new point and k of the closest previously labeled observations. Of the k neighbors, the majority class is chosen to be the label for the test point. For example, with k=5, we look at the features of a new person whose risk has not been identified yet and look at 5 people with the features closest matching this new person out of the training set. If 3 of them are high-risk, and two are low-risk, the new person is identified as high risk with 3 votes to 2. For even numbers of k such as 6 where there might be ties, we weight the votes by proximity. So, with respect to our PCA features, we are comparing people who closely have similar sleep characteristics, data usage and so on. We found k=2 to perform best in our scenario, likely due to the low sample size, as well as high variability amongst users. We used a Euclidean distance metric and enforced that each feature have equal distance weighting (uniform weights).

Discussion

Principal Results

Results from this feasibility study indicate that, although not a perfect predictor, the K-Nearest Neighbors model is suitable for this study, as it has shown ability to separate users deemed as at risk from the Columbia Risk Assessment (C-SSRS) to those not deemed at risk at an average rate beyond just randomly guessing (i.e. at an average rate 15% beyond randomly guessing the majority to be at low risk). These are early indications that it is possible to predict risk using the data collected in this feasibility study, using the KNN algorithm. The data used to inform this included a user's sleeping activity, step activity, self-reported mood, journaling thoughts, and activity levels with respect to the phone app.

This is a crucial first step in automatic risk assessment as we have verification of signal from data that were collected directly from smartphone interactions. This is also promising as we are working with a relatively small data set from a machine learning perspective. This is the base for future phases of this study where we will be looking to test the model on additional users of mental health services for further testing of concept and generalizability.

The implications of this feasibility study are highly significant for building capacity for suicide risk prediction (future risk) and/or detection (real time/current risk). With a low proportion of suicide attempters/completers who actually access mental health services [61], it is essential to develop and test non-clinical means of assessing risk. Given the dynamic nature of suicide ideation and suicide risk, new methods are needed to track suicide risk in real time [62], together with a better understanding of the ways in which people communicate or express their suicidality [25]. Mobile applications (apps) could be better suited to help prevent suicide by offering support in situ and at the time of crisis [63].

While prior studies have utilized electronic health care record data to create an actuarial model of suicide risk [30,34,36,38,65-67], or focused on a single aspect of user input such as language [29-30,37], this study adds to the literature by introducing inclusion of external, user generated input and smartphone data and combining it with clinical data. Our study adds to evidence that report on the use of external, non-clinical data to predict suicidality. The results are promising, even though we used 'basic', 'simpler', 'routine' biometrics (collected via iphone and Fitbit), compared to data used in previous research; similar studies aiming to predict mental state

(short term) have used multiple (self-report) measurements and a wide range of bio sensors [12,15,76].

Strengths, Limitations and Further Testing

We recognize that our study is limited by the short follow up period of up to a week, thus future iterations would need to extend to a longer period of study to explore the time sensitivity of model predictions over varying time windows (e.g. predicting current risk vs one week out). Short-term risk prediction is difficult as any inference is based on a small amount/limited data which means that meaningful signals are lost due to noise from highly variable behaviours [13]. There is promise in improvement as the amount of data available for training and testing increases. Previous research and machine learning literature [68-71] points to expected improvement in performance and reliability in test results as sample size grows, particularly in this classification setting. We expect roughly doubling the sample size would achieve more practical results to where the possibility of implementation would be appropriate.

Although the results from this feasibility study have been promising in producing a signal, in terms of operationalizing risk for suicide, future steps would be moving beyond survey generated risk scores. Before taking that leap, the intermediate step would be to further validate the algorithmic results by collecting additional, more substantial test data. Where the experiment excels is in the data sources are diverse rather than strictly clinical and allows for natural extension to outpatient settings. Also, given the probabilistic nature of the algorithm, there will naturally continue to be a tradeoff between false positives and false negatives as the model improves and hence medical/human attention in decision making will remain critical. We propose the algorithmic approach provides a supplement and an additional facet to clinical judgement.

Therefore, having achieved a signal from the data for risk in Phase I, Phase II (proof of concept) will involve collecting more data to not only see if modeling improves, but also to test other models such as predicting the risk score trajectory. Enforcing a minimum of two C-SSRS assessments, we can try to model changes in risk. We also intend to experiment with more

features, particularly those involved in text mining as most journaling features were relatively surface level. Moreover, we will be aiming to look at prediction stability over time as this prediction was made within a couple days from usage to assessment.

Our final aim is to form our own standard so as to break away from dependency on the C-SSRS, as we look to go beyond information gathered in a formal survey that depends solely on human judgment. Further research will enable us to test viability of automation and machine learning to identify suicide risk by comparing predictions of risk to eventual outcomes, as well as testing out the model in different settings and populations (e.g. community).

We would also like to point out that, while mobile phones and applications are ubiquitous nowadays and have the potential to be an efficient and cost-effective approach to addressing mental health problems [72], this study indicated that there are certain costs that limit the widespread adoption of health apps within mental health services (weather inpatient and community settings). These are related to access to smartphones, connectivity, updating and maintenance of technology. The premise for this study was that, in line with the UK population statistics indicating that approximately 95 percent of households own a mobile phone [73], of which a high proportion are smartphones, participants would have access to and use their own smartphones for the study. Following initial scoping, the authors realised that only a small proportion of inpatients had access to a smartphone. In addition, the SWiM app was configured (in its current testing form) to operate only with IOS products, i.e. an iPhone. We cannot confirm the extent to which given participants study iPhones might have affected the results, this is something that needs to be further explored. We can, however, highlight that participants were enthusiastic about using FitBit wearables and the FitBit app on the phone, which may or may not have not encouraged them to use the SWiM app as well.

Conclusion

Although in its early stages, research in this area suggests that using smartphones to enquire about suicidal behavior can be a valuable approach and not a risk factor for increasing suicidal ideation [12]. Given the heavy reliance on smartphones and mobile devices around the world, this readily available source of data is an important and highly underutilized source that has

good potential to improve mental health risk prediction and prevention and advance global mental health.

However, while full automation and independence of clinical judgement or input would be a worthy development for those individuals who are less likely to access specialist mental health services, and for providing a timely response in a crisis situation, we need to acknowledge the ethical and legal implications of such advances in the field of psychiatry [73-74]. The use of machine learning in suicide prediction needs a strong evidence base across different settings, populations, suicidal behaviours and datasets, before considering a full integration in healthcare settings. For the time being, if proven accurate and scalable, machine learning algorithms for suicide risk detection are likely to complement rather than replace clinical judgement [73]. While smartphones provide us with the opportunities to gather data on real time dynamic risk factors for suicide behaviour which would be almost impossible to monitor on discharge (from mental health settings), more research is needed to validate the utility of risk markers for suicide behaviour and confirm a safe and clinically effective way to use these data to inform practice [13]. More work is needed before we can achieve safe and effective integration within mental health settings, while remaining attentive to key ethical implications. An interesting ethical dimension is related to the use of the k-nearest neighbors (KNN) algorithm that requires continued access to the pooled data of (at least a subset of) multiple participants in order to subsequently label new cases. While testing this in a controlled setting with inpatients who have provided consent for the use of their data might be straightforward, it is uncertain whether service users in the community would accept to have their suicidal trajectory data shared for this purpose or how mental health services would be able to bridge the gap. Furthermore, to achieve high accuracy in terms of short term risk prediction, a wide variety of data from multiple sources will need to be collected, with data integration as a key component [13]. We therefore expect multiple data governance, privacy and IP issues at stake.

Multimedia Appendices

Multimedia Appendix 1: Sample training data

References

1. Capodanno AE, Targum SD. Assessment of suicide risk: some limitations in the prediction of infrequent events. Journal of psychosocial nursing and mental health services. 1983;21(5):11-4. doi: 10.3928/0279-3695-19830501-03
2. Franklin JC, Ribeiro JD, Fox KR, Bentley KH. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. Psychological bulletin. 2017;143(2):187-232. PMID: 27841450. doi: 10.1037/bul0000084
3. Haney EM, O'Neil ME, Carson S, Low A, Peterson K, Denneson LM, et al. Suicide Risk Factors and Risk Assessment Tools: A Systematic Review. Portland: Evidence-based Synthesis Program (ESP) Center; 2012. PMID: 22574340
4. Ribeiro JD, Franklin JC, Fox KR, Bentley KH, Kleiman EM, Chang BP, et al. Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. Psychological medicine. 2016;46(2):225-36. doi: 10.1017/S0033291715001804
5. Turecki G, Brent DA. Suicide and suicidal behaviour. Lancet (London, England). 2016;387(10024):1227-39. PMID: 26385066. doi: 10.1016/S0140-6736(15)00234-2
6. Goldsmith SK, Pellmar TC, Kleinman, AM, Bunney, WE (Eds.) (2002). Reducing Suicide: A National Imperative. Washington, DC: Institute of Medicine
7. McAuliffe CM. Suicidal ideation as an articulation of intent: a focus for suicide prevention? Arch. Suicide Res. 2002;6: 325–338. doi: 10.1080/13811110214524
8. Wolk-Wasserman D. Suicidal communication of persons attempting suicide and responses of significant others. Acta Psychiatr. Scand. 1986;73: 481–499. PMID: 3751655. doi: 10.1111/j.1600-0447.1986.tb02715.x
9. Wasserman D, Tran Thi Thanh H, Pham Thi Minh D, Goldstein M, Nordenskiöld A, Wasserman C. Suicidal process, suicidal communication and psychosocial situation of young suicide attempters in a rural Vietnamese community. World Psychiatry. 2008;7: 47–53. PMID: 18458785. doi: 10.1002/j.2051-5545.2008.tb00152.x
10. O'Dea B, Wan S, Batterhamc P J, Calear AL, Paris C, Christensen H. Detecting suicidality on Twitter. Internet Interventions. 2015;2: 183–188. doi: 10.1016/j.invent.2015.03.005

11. Czyz EK, Horwitz AG, Eisenberg D, Kramer A, King CA. Self-reported barriers to professional help seeking among college students at elevated risk for suicide. Journal of American college health: J of ACH. 2013;61(7):398–406. PMID: 24010494. doi: 10.1080/07448481.2013.820731

12. Kleiman EM, Turner BJ, Fedor S, Beale EE, Picard RW, Huffman JC, Nock MK. Digital phenotyping of suicidal thoughts. Depression and Anxiety. 2018;35(7), 601-608. PMID: 29637663. doi: 10.1002/da.22730

13. Torous J, Larsen ME, Depp C, Cosco TD, Barnett I, Nock MK, Firth J. Smartphones, Sensors, and Machine Learning to Advance Real-Time Prediction and Interventions for Suicide Prevention: a Review of Current Progress and Next Steps. Current Psychiatry Reports. 2018;20(7), 51. PMID: 29956120. doi: 10.1007/s11920-018-0914-y

14. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol. 2008;4:1–32. PMID: 18509902. doi: 10.1146/annurev.clinpsy.3.022806.091415

15. Kleiman EM, Nock, MK. Real-time assessment of suicidal thoughts and behaviors. Current Opinion in Psychology. 2018;22, 33–37. oi: 10.1016/j.copsyc.2017.07.026

16. Bidargaddi N, Musiat P, Makinen V-P, Ermes M, Schrader G, Licinio J. Digital footprints: Facilitating large-scale environmental psychiatric research in naturalistic settings through data from everyday technologies. Molecular Psychiatry. 2017;22(2), 164–169. PMID: 27922603. doi: 10.1038/mp.2016.224

17. Onnela J-P, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology. 2016;41(7), 1691–1696. PMID: 26818126. doi: 10.1038/npp.2016.7

18. Torous J. Onnela J-P, Keshavan M. New dimensions and new tools to realize the potential of RDoC: Digital phenotyping via smartphones and connected devices. Translational Psychiatry. 2017;7(3), e1053. PMID: 28267146. doi: 10.1038/tp.2017.25

19. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res. 2009;11(1):e11. PMID: 19329408. doi: 10.2196/jmir.1157

20. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. Am J Prev Med. 2011;40(5 Suppl 2):S154-S158. PMID: 21521589. doi: 10.1016/j.amepre.2011.02.006

21. Tovar D, Cornejo E, Xanthopoulos P, Guarracino MR, Pardalos PM. Data mining in psychiatric research. Methods in Molecular Biology (Clifton, NJ). 2012;829: 593–603

22. Ruder TD, Hatch GM, Ampanozi G, Thali MJ, Fischer N. Suicide announcement on Facebook. Crisis. 2011;32(5): 280–282. PMID: 21940257. doi: 10.1027/0227-5910/a000086

23. Luxton DD, June JD, Fairall JM. Social media and suicide: a public health perspective. Am. J. Public Health. 2012;102 (2): S195–S200. PMID: 22401525. doi: 10.2105/AJPH.2011.300608

24. Jashinsky J, Burton SH, Hanson CL, West J, Giraud-Carrier C, Barnes MD, Argyle T. Tracking suicide risk factors through Twitter in the US. Crisis. 2013;35: 51–59. PMID: 24121153. doi: 10.1027/0227-5910/a000234

25. O'dea B, Larsen ME, Batterham PJ, Calear AL, Christensen H. A linguistic analysis of suicideralated Twitter posts. Crisis. 2017;38(5):319-329. PMID: 28228065. doi: 10.1027/0227-5910/a000443

26. Burnap P, Colombo G, Amery R, Hodorog A, Scourfield J. Multiclass machine classification of suicide-related communication on Twitter. Online Soc Netw Media. 2017;2:32–44. PMID: 29278258. doi: 10.1016/j.osnem.2017.08.001

27. Guan L, Hao B, Cheng Q, Yip PS, Zhu T.Identifying Chinese Microblog Users With High Suicide Probability Using Internet-Based Profile and Linguistic Features: Classification Model. JMIR Mental Health. 2015;2(2): e17. PMID: 26543921. doi: 10.2196/mental.4227

28. Beiwinkel T, Kindermann S, Maier A, Kerl C, Moock J, Barbian G, Rössler W. Using Smartphones to Monitor Bipolar Disorder Symptoms: A Pilot Study. JMIR Ment Health. 2016;3(1):e2. PMID: 26740354. doi: 10.2196/mental.4560

29. Pestian JP, Sorter M, Connolly M, Cohen KB, McCullum Smith C, Gee JT, Morency L-P, Scherer S, Rohlfsa L. Machine Learning Approach to Identifying the Thought Markers of Suicidal Subjects: A Prospective Multicenter Trial. Suicide and Life-Threatening Behavior. 2016;1-9. PMID: 27813129. doi: 10.1111/sltb.12312

30. Kessler RC, Hwang I, Hoffmire CA,McCarthy JF, Petukhova MV, Rosellini AJ, Sampson NA, Schneider AL, Bradley PA, Katz IR, Thompson C. Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration. Int J Methods Psychiatr Res. 2017;26(3). PMID: 28675617. doi: 10.1002/mpr.1575

31. Morales S, Barros J, Echávarri O, García F, Osses A, Moya C, et al. Acute mental discomfort associated with suicide behavior in a clinical sample of patients with affective disorders: ascertaining critical variables using artificial intelligence tools. Front Psych. 2017;8:7. PMID: 28210230. doi: 10.3389/fpsyt.2017.00007

32. Passos IC, Mwangi B, Cao B, Hamilton JE, Wu M-J, Zhang XY, Zunta-Soares GB, Quevedo J, Kauer-Sant'Anna M, Kapczinski F, Soares JC. Identifying a clinical signature of suicidality among patients with mood disorders: A pilot study using a machine learning approach. Journal of Affective Disorders. 2016;193: 109–116. PMID: 26773901. doi: 10.1016/j.jad.2015.12.066

33. Delgado-Gomez D, Blasco-Fontecilla H, Sukno F, Ramos-Plasencia MS, Baca-Garcia E. Suicide attempters classification: Toward predictive models of suicidal behavior. Neurocomputing: An International Journal. 2012;92:3-8. doi: 10.1016/j.neucom.2011.08.033

34. Simon GE, Johnson E, Lawrence JM, Rossom RC, Ahmedani B, Lynch FL, Beck A, Waitzfelder B, Ziebell R, Penfold RB, Shortreed SM. Predicting Suicide Attempts and Suicide Deaths Following Outpatient Visits Using Electronic Health Records.

American Journal of Psychiatry. 2018;175(10):951-60. doi: 10.1176/appi.ajp.2018.17101167

35. Pestian JP, Sorter M, Connolly B, Bretonnel Cohen K, McCullumsmith C, Gee JT, Morency LP, Scherer S, Rohlfs L, STM Research Group. A Machine Learning Approach to Identifying the Thought Markers of Suicidal Subjects: A Prospective Multicenter Trial. Suicide & life-threatening behavior. 2017;47(1):112-21. PMID: 27813129 doi: 10.1111/sltb.12312
36. Poulin C, Shiner B, Thompson P, Vepstas L, Young-Xu Y, Goertzel B, Watts B, Flashman L, McAllister T. Predicting the risk of suicide by analyzing the text of clinical notes. PLoS One. 2014;9(1):e85733. PMID: 24489669. doi: 10.1371/journal.pone.0085733
37. Thompson P, Bryan C, Poulin C, editors. Predicting military and veteran suicide risk: Cultural aspects. Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality; 2014; Baltimore, Maryland, USA: Association for Computational Linguistics
38. Walsh CG, Ribeiro JD, Franklin JC. Predicting Risk of Suicide Attempts Over Time Through Machine Learning. Clinical Psychological Science. 2017;5(3):457-69. doi: 10.1177/2167702617691560
39. Posner K, Brown GK, Stanley B, Brent DA, Yershova KV, Oquendo MA, Currier GW, Melvin GA, Greenhill L, Shen S, Mann JJ. The Columbia-Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am J Psychiatry. 2011;168(12):1266-77. PMID: 22193671. doi: 10.1176/appi.ajp.2011.10111704
40. FDA. Guidance for Industry. Suicidal Ideation and Behavior: Prospective Assessment of Occurrence in Clinical Trials. Draft Guidance. U.S. Department of Health and Human Services Food and Drug Administration (FDA). Center for Drug Evaluation and Research (CDER). August 2012. Clinical/Medical Revision 1. Available to download: http://www.fda.gov/downloads/Drugs/.../Guidances/UCM225130.pdf. Accessed on 4th January 2017

41. Ribeiro JD, Pease JL, Gutierrez PM, Silva C, Bernert RA, Rudd MD, Joiner TE Jr. Sleep problems outperform depression and hopelessness as cross-sectional and longitudinal predictors of suicidal ideation and behavior in young adults in the military. Journal of affective disorders. 2012;136(3):743-50. PMID: 22032872. doi: 10.1016/j.jad.2011.09.049

42. Hutto CJ, Gilbert E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media; 2014; Ann Arbor, MI: Association for the Advancement of Artificial Intelligence

43. Guo X, Yin Y, Dong C, Yang G, Zhou G, editors. On the Class Imbalance Problem. Proceedings of the 2008 Fourth International Conference on Natural Computation; 2008; Jinan: IEEE Computer Society

44. Kohavi, R. A Study pf Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Appears in the International Joint Conference on Artificial Intelligence (IJCAI), 1995

45. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second ed: Springer Series in Statistics; 2017. ISBN 978-0-387-84858-7

46. Lever J, Krzywinski M, Altman N. Principal component analysis. Nature Methods. 2017;14:641. doi: 10.1038/nmeth.4346

47. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32. doi: 10.1023/A:1010933404324

48. Louppe G, Wehenkel L, Sutera A, Geurts P, editors. Understanding variable importances in forests of randomized trees. Proceedings of the 26th International Conference on Neural Information Processing Systems; 2013; Lake Tahoe, Nevada

49. Biau G. Analysis of a Random Forests Model. Journal of Machine Learning Research. 2012;13:1063-95

50. Hearst MA. Support Vector Machines. IEEE Intelligent Systems. 1998;13(4):18-28. doi: 10.1109/5254.708428

51. Cho CH, Lee T, Kim MG, In HP, Kim L, HJ L. Mood Prediction of Patients With Mood Disorders by Machine Learning Using Passive Digital Phenotypes Based on the Circadian Rhythm: Prospective Observational Cohort Study. J Med Internet Res. 2019 Apr 17;21(4):e11029. PMID: 30994461. doi: 10.2196/11029

52. Krystal AD. PSYCHIATRIC DISORDERS AND SLEEP. Neurol Clin 2012;30(4):1389-413. PMID: 23099143. doi: 10.1016/j.ncl.2012.08.018

53. Lampinen P, Heikkinen RL, Kauppinen M, Heikkinen E. Activity as a predictor of mental well-being among older adults. Aging & mental health. 2006;10(5):454-66. PMID: 16938681. doi: 10.1080/13607860600640962

54. Sheaves B, Porcheret K, Tsanas A, Espie CA, Foster RG, Freeman D, Harrison PJ, Wulff K, Goodwin GM. Insomnia, Nightmares, and Chronotype as Markers of Risk for Severe Mental Illness: Results from a Student Population. Sleep. 2016;39(1):173-81. PMID: 26350467. doi: 10.5665/sleep.5342

55. Domingos P, editor. A Unified Bias-Variance Decomposition for Zero-One and Squared Loss. Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence; 2000: AAAI Press

56. Geman S, Bienenstock E, Doursat R. Neural Networks and the Bias/Variance Dilemma. Neural Computation. 1992;4(1):1-58. doi: 10.1162/neco.1992.4.1.1

57. Belkin M, Hsu D, Ma S, Mandal S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proceedings of the National Academy of Sciences Aug 2019, 116 (32) 15849-15854; DOI:10.1073/pnas.1903070116

58. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967;13(1):21-7. doi: 10.1109/TIT.1967.1053964

59. Alpaydin E. Combined 5×2 cv F test for comparing supervised classification learning algorithms. Neural computation. 1999:11(8), 1885-1892

60. Dietterich TG. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Comput. 1998;10:1895–1923

61. Healthcare Quality Improvement Partnership. National Confidential Inquiry into Suicide and Homicide by People with Mental Illness. Annual Report. 2017

62. Braithwaite SR, Giraud-Carrier C, West J, Barnes MD, Hanson CL. Validating Machine Learning Algorithms for Twitter Data Against Established Measures of Suicidality. JMIR mental health. 2016;3(2):e21. PMID: 27185366. doi: 10.2196/mental.4822

63. Larsen ME, Nicholas J, Christensen H. A Systematic Assessment of Smartphone Tools for Suicide Prevention. PLoS One. 2016;11(4):e0152285. PMID: 27073900. doi: 10.1371/journal.pone.0152285

64. Walsh CG, Ribeiro JD, Franklin JC. Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. Journal of child psychology and psychiatry, and allied disciplines. 2018;59(12):1261-70. doi: 10.1111/jcpp.12916

65. Barak-Corren Y, Castro VM, Javitt S, Hoffnagle AG, Dai Y, Perlis RH, Nock MK, Smoller JW, Reis BY. Predicting Suicidal Behavior From Longitudinal Electronic Health Records. Am J Psychiatry. 2017;174(2):154-62. PMID: 27609239. doi: 10.1176/appi.ajp.2016.16010077

66. Kessler RC, Warner CH, Ivany C, Petukhova MV, Rose S, Bromet EJ, Brown M 3rd, Cai T, Colpe LJ, Cox KL, Fullerton CS, Gilman SE, Gruber MJ, Heeringa SG, Lewandowski-Romps L, Li J, Millikan-Bell AM, Naifeh JA, Nock MK, Rosellini AJ, Sampson NA, Schoenbaum M, Stein MB, Wessely S, Zaslavsky AM, Ursano RJ; Army STARRS Collaborators. Predicting suicides after psychiatric hospitalization in US Army soldiers: the Army Study To Assess Risk and rEsilience in Servicemembers (Army STARRS). JAMA psychiatry. 2015;72(1):49-57. PMID: 25390793. doi: 10.1001/jamapsychiatry.2014.1754

67. Kessler RC, van Loo HM, Wardenaar KJ, Bossarte RM, Brenner LA, Cai T, Ebert DD, Hwang I, Li J, de Jonge P, Nierenberg AA, Petukhova MV, Rosellini AJ, Sampson NA, Schoevers RA, Wilcox MA, Zaslavsky AM. Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. Molecular psychiatry. 2016;21(10):1366-71. PMID: 26728563. doi: 10.1038/mp.2015.198

68. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. BMC Medical Informatics and Decision Making. 2012; 12:8. PMID: 22336388. doi: 10.1186/1472-6947-12-8

69. Popovici V, Chen W, Gallas BG, Hatzis C, Shi W, Samuelson FW, Nikolsky Y, Tsyganova M, Ishkin A, Nikolskaya T, Hess KR, Valero V, Booser D, Delorenzi M, Hortobagyi GN, Shi L, Symmans WF, Pusztai L. Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. Breast cancer research :BCR. 2010;12(1), R5. PMID: 21092148. doi: 10.1186/gm202

70. Beleites C, Neugebauer U, Bocklitz T, Krafft C, Popp J. Sample Size Planning for Classification Models. Analytica Chimica Acta. 2013;760: 5 – 33. PMID: 23265730. doi: 10.1016/j.aca.2012.11.007

71. Sordo M, Zeng Q. On Sample Size and Classification Accuracy: A Performance Comparison. In: Oliveira J.L., Maojo V., Martín-Sánchez F., Pereira A.S. (eds) Biological and Medical Data Analysis. ISBMDA 2005. Lecture Notes in Computer Science, vol 3745. Springer, Berlin, Heidelberg

72. Olff M. Mobile mental health: a challenging research agenda. European Journal of Psychotraumatology. 2015;6(1), 27882. doi:10.3402/ejpt.v6.27882

73. Office of National Statistics. Percentage of households with mobile phones in the United Kingdom (UK) from 1996 to 2018 (2019). Available at: https://www.statista.com/statistics/289167/mobile-phone-penetration-in-the-uk/. Accessed 2nd December 2019

74. Tiffin PA, Paton LW. Rise of the machines? Machine learning approaches and mental health: opportunities and challenges. British Journal of Psychiatry. 2018;213(3):509-10. PMID: 30113285. doi: 10.1192/bjp.2018.105

75. Linthicum KP, Schafer KM, Ribeiro JD. Machine learning in suicide science: Applications and ethics. Behavioral sciences & the law. 2019; 37(3): 214-222. doi: 10.1002/bsl.2392

76. Alam MGR, Cho EJ, Huh E-N, Hong CS. Cloud Based Mental State Monitoring System for Suicide Risk Reconnaissance Using Wearable Bio-sensors. Proceedings of the 8th

International Conference on Ubiquitous Information Management and Communication. January 2014. Article No.: 56. Pages 1–6. doi: 10.1145/2557977.2558020

Declaration of interest

Dr. Haines, Mr Chahal, Miss Bruen, Miss Wall, and Mr Sadashiv have nothing to disclose.

Dr. Khan reports grants from Stanford University School of Medicine, during the conduct of the study; personal fees from The Risk Authority in 2016 prior to the study, outside the submitted work. Dr. Fearnley reports being a member of the Board of Managers for Innovence Augmented Intelligence Medical Systems Psychiatry, an LLC between Mersey Care NHS Foundation Trust and The Risk Authority, Stanford. This has overseen the development of the technology that is undergoing evaluation in this research study.

Acknowledgements

Funding: this work was supported by the Department of Health, Global Digital Exemplar (ref: CENT/DIGEX/RW4/2017-10-16/A).

Author contribution: Dr. Haines has contributed to the design of the study; set up and management of the study; has drafted sections of the manuscript and revised it critically. Mr Chahal has conducted all the data analysis and interpretation, drafted sections of the manuscript, as well as contributing to the overall revision. Miss Bruen and Miss Wall have collected all the data for the work and have critically revised the manuscript. Dr Khan has contributed to the design of the study, has drafted sections of the manuscript and has critically revised it. Mr Sadashiv was responsible for managing the design and development of the

software application, product analytics and delivery, as well as revising the manuscript before submission. Dr Fearnley was the CI for this project, having contributed to the design and oversight of the project, as well as drafting sections of the manuscript and critically reviewing it. Other acknowledgements: The authors would like to acknowledge Abhijith Nagaraja's contribution. He was the lead architect and engineer to design and deploy the iOS applications used to collect data for study participants. Abhijith was also responsible to design and develop the backend software to ensure that data is captured and stored as per privacy policy standards requirements.